BISS Biodiversity Information Science and Standards

Biodiversity Information Science and Standards 4: e59088
doi: 10.3897/biss.4.59088

OPEN ACCESS

Conference Abstract

Enabling Machines to Integrate Biodiversity Data with Evolutionary Knowledge

Gaurav Vaidya[‡], Hilmar Lapp[§], Nico Cellinese[|]

‡ University of North Carolina, Chapel Hill, United States of America
§ Duke University, Durham, NC, United States of America
| University of Florida, Gainesville, United States of America

Corresponding author: Gaurav Vaidya (gaurav@renci.org)

Received: 29 Sep 2020 | Published: 02 Oct 2020

Citation: Vaidya G, Lapp H, Cellinese N (2020) Enabling Machines to Integrate Biodiversity Data with Evolutionary Knowledge. Biodiversity Information Science and Standards 4: e59088.
https://doi.org/10.3897/biss.4.59088

Abstract

Most biological data and knowledge are directly or indirectly linked to biological taxa via taxon names. Using taxon names is one of the most fundamental and ubiquitous ways in which a wide range of biological data are integrated, aggregated, and indexed, from genomic and microbial diversity to macro-ecological data. To this day, the names used, as well as most methods and resources developed for this purpose, are drawn from Linnaean nomenclature. This leads to numerous problems when applied to data-intensive science that depends on computation to take full advantage of the vast – and rapidly increasing – amount of available digital biodiversity data. The theoretical and practical complexities of reconciling taxon names and concepts has plagued the systematics community for decades and now more than ever before, Linnaean names based in Linnaean taxonomy, by far the most prevalent means of linking data to taxa, are unfit for the age of computation-driven data science, due to fundamental theoretical and practical shortfalls that cannot be cured.

We propose an alternate approach based on the use of phylogenetic clade definitions, which is a well-developed method for unambiguously defining the semantics of a clade concept in terms of shared evolutionary ancestry (de Queiroz and Gauthier 1990, de Queiroz and Gauthier 1994). These semantics allow locating the defined clade on any phylogeny, or showing that a clade is inconsistent with the topology of a given phylogeny

and hence cannot be present on it at all. We have built a workflow for defining phylogenetic clade definitions in terms of shared ancestor and excluded lineage properties, and locating these definitions on any input phylogeny. Once these definitions have been located, we can use the list of species found within that clade on that phylogeny in order to aggregate occurrence data from the Global Biodiversity Information Facility (GBIF). Thus, our approach uses clade definitions with machine-understandable semantics to programmatically and reproducibly aggregate biodiversity data by higher-level taxonomic concepts. This approach has several advantages over the use of taxonomic hierarchies:

1. Unlike taxa, the semantics of clade definitions can be expressed in unambiguous, machine-understandable and reproducible terms and language.

2. The resolution of a given clade definition will depend on the phylogeny being used. Thus, if the phylogeny of groups of interest is updated in light of new evolutionary knowledge, the clade definition can be applied to the new phylogeny to obtain an updated list of clade members consistent with the updated evolutionary knowledge.

3. Machine reproducibility of analyses is possible simply by archiving the machine-readable representations of the clade definition and the phylogeny being used.

Clade definitions can be created by biologists as needed or can be reused from those published in peer-reviewed journals. In addition, nearly 300 peer-reviewed clade definitions were recently published as part of the Phylonym volume of the PhyloCode (de Queiroz et al. 2020) and are now available on the Regnum website. As part of the Phyloreferencing Project, we digitize this collection as a machine-readable ontology, where each clade is represented as a class defined by logical conjunctions for class membership, corresponding to a set of necessary and sufficient conditions of shared or divergent evolutionary ancestry. We call these classes phyloreferences, and have created a fully automated workflow for digitizing the Regnum database content into an OWL ontology (W3C OWL Working Group 2012) that we call the Clade Ontology. This ontology includes reference phylogenies and additional metadata about the verbatim clade definitions. Once complete, the Clade Ontology will include all clade definitions from RegNum, both those included in Phylonym after passing peer-review, and those contributed by the community, whether or not under the PhyloCode nomenclature. As an openly available community resource, this will allow researchers to use them to aggregate biodiversity data for comparative biology with grouping semantics that are transparent, machine-processable, and reproducible.

In our presentation, we will demonstrate the use of phyloreferences to locate clades on the Open Tree of Life synthetic tree (Hinchliff et al. 2015), to retrieve lists of species in each clade, and to use them to find and aggregate occurrence records in GBIF. We will also describe the workflow we are currently using to build and test the Clade Ontology, and describe our plans for publishing this resource. Finally, we will discuss the advantages and disadvantages of this approach as compared to taxonomic checklists.

Keywords

phylogenetics, clade definitions, ontologies, ontology development, phyloreferences

Presenting author

Gaurav Vaidya

Presented at

TDWG 2020

Funding program

The Phyloreferencing project is funded by the US National Science Foundation through collaborative grants DBI-1458484 and DBI-1458604 to Hilmar Lapp (Duke University) and Nico Cellinese (University of Florida), respectively. The proposal text is available online (Cellinese and Lapp 2015).

References

- Cellinese N, Lapp H (2015) An ontology-based system for querying life in a post-taxonomic age. Figshare. https://doi.org/10.6084/M9.FIGSHARE.1401984
- de Queiroz K, Gauthier J (1990) Phylogeny as a central principle in taxonomy: Phylogenetic definitions of taxon names. Systematic Biology 39 (4): 307-322. https://doi.org/10.2307/2992353
- de Queiroz K, Gauthier J (1994) Toward a phylogenetic system of biological nomenclature. Trends in Ecology & Evolution 9 (1): 27-31. https://doi.org/10.1016/0169-5347(94)90231-3
- de Queiroz K, Cantino P, Gauthier J (Eds) (2020) Phylonyms: A companion to the PhyloCode. 1st Edition. CRC Press, 1324 pp. URL: https://www.routledge.com/Phylonyms-A-Companion-to-the-PhyloCode/Queiroz-Cantino-Gauthier/p/book/9781138332935 [ISBN 9781138332935]
- Hinchliff C, Smith S, Allman J, Burleigh JG, Chaudhary R, Coghill L, Crandall K, Deng J, Drew B, Gazis R, Gude K, Hibbett D, Katz L, Laughinghouse IV HD, McTavish EJ, Midford P, Owen C, Ree R, Rees J, Soltis D, Williams T, Cranston K (2015) Synthesis of phylogeny and taxonomy into a comprehensive tree of life. Proceedings of the National Academy of Sciences 112 (41): 12764-12769. https://doi.org/10.1073/pnas.1423041112
- W3C OWL Working Group (2012) OWL 2 Web Ontology Language Document Overview (Second Edition). https://www.w3.org/TR/2012/REC-owl2-overview-20121211/. Accessed on: 2020-8-12.